Research Report

Different Methods of Adjusting for Form Difficulty Under the Rasch Model: Impact on Consistency of Assessment Results

ETS RR–19-08

Venessa F. Manna
Lixiong Gu

December 2019

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Different Methods of Adjusting for Form Difficulty Under the Rasch Model: Impact on Consistency of Assessment Results

Venessa F. Manna & Lixiong Gu

Educational Testing Service, Princeton, NJ

When using the Rasch model, equating with a nonequivalent groups anchor test design is commonly achieved by adjustment of new form item difficulty using an additive equating constant. Using simulated 5-year data, this report compares 4 approaches to calculating the equating constants and the subsequent impact on equating results. The 4 approaches are mean difference, mean difference with outlier removal using the 0.3 logit rule, mean difference with robust *z* statistic, and the information-weighted mean difference. Factors studied included sample size, anchor test length, percentage of anchor items displaying outlier behavior, and the distribution of test item difficulty relative to examine ability. The results indicated that the mean difference and information-weighted mean difference methods performed similarly across all conditions. In addition, with larger sample sizes, the mean difference with 0.3 logit method performed similarly to these 2 methods. The mean difference with robust *z* method performed most differently from the other three methods of calculating the equating constant. This method removed a large percentage of the anchor items compared to the mean difference with 0.3 logit method but seemed to produce the most stable trend in performance classification across the 5 years, particularly when sample sizes were large.

Keywords Rasch model; equating; item response theory; outliers; calibration

doi:10.1002/ets2.12244

With ever-increasing reliance on large-scale assessment for decision-making at the institutional and individual levels, the accuracy and consistency of test results are fundamentally important. For example, in many K – 12 accountability systems, results of annually administered large-scale assessments may be used to monitor academic achievement and improvement or growth in student performance over time. At the individual level, scores may be used for admissions to institutions of higher learning, eligibility for scholarships or special programs, demonstration of language proficiency in the workplace, and study-abroad programs.

For security reasons, large-scale assessment programs often use alternate test forms that are administered at different points in time. These forms are usually constructed to the same content and statistical specifications, and test-equating procedures are used to establish score comparability across forms. One of the most widely used equating designs is the nonequivalent groups anchor test (NEAT) design (Holland & Dorans, 2006). In the NEAT design, a set of anchor items (also called common items) is embedded in the alternate forms (new and reference) to be equated. It is the performance of examinees on these anchor items that is used to adjust the difficulty of alternate forms of a test.

Item response theory (IRT) is often used to establish and maintain score scales for alternative forms of a test. One of the most important features of IRT is the invariance of item parameters. Theoretically, the item parameters are invariant across samples of examinees from the same population, except for a linear difference between proficiency scales due to scale indeterminacy in IRT modeling (Lord, 1980). IRT equating, like many other IRT applications, relies on the invariance property. With the NEAT design, the new and reference forms are separately calibrated using different samples, and the item parameters on the new form can be placed onto the scale of the reference form by a simple linear transformation based on the performance of the anchor items (Cook, Eignor, & Taft, 1998; Eignor & Cook, 1983). For those testing programs that employ IRT calibration and equating, the Rasch model is popular, in part, due to the easily understood one-to-one relationship between examinees' raw scores and scale scores. For this model, item parameters on the new form can be easily transformed to the scale of the reference form using an additive constant calculated as the mean difference of anchor item

Corresponding author: V. F. Manna, E-mail: vmanna@ets.org

difficulty on the new and reference forms (commonly called the mean difference method). Therefore estimation of the equating function is closely linked to the precision of the item parameter estimates. Moreover, a key assumption is that the measurement properties of the anchor items are stable across the test forms (Dorans, 1986; Hanson & Feinstein, 1997; Wainer, 1999). In reality, estimates of anchor item parameters do vary, and it is a common practice to review the parameter estimates of the anchor items for consistency between the test forms.

Anchor items can show differential performance due to item ordering and context effects. Yen (1980) found that reading comprehension items generally became more difficult when they were placed later in the test than when they were placed at the beginning. Results consistent with those of Yen (1980) were reported in studies by Eignor and Cook (1983) and Kingston and Dorans (1984). Editorial changes to anchor items (Cassels & Johnstone, 1984), changes in the ordering of multiple-choice options (Cizek, 1994; Tollefson, 1987), and uncontrolled rater effects (Fitzpatrick, Ercikan, Yen, & Ferrara, 1998) are additional factors that can cause the anchor items to behave differently across forms and administrations. Kolen and Brennan (2004, pp. 307–309) provided a comprehensive list of nonachievement factors that can change the measurement properties of anchor items. In addition, changes in instructional emphasis from 1 year to the next can result in differential relative performance for different items and item types (Bock, Muraki, & Pfeiffenberger, 1988; Masters, 1988; A. D. Miller & Linn, 1988).

Various studies have shown that anchor items that exhibit large differential performance and are thus considered outliers can have an effect on the results of IRT-based equating (Hanson & Feinstein, 1997; Hu, Rogers, & Vukmirovic, 2008; Karkee & Choi, 2005; Michaelides, 2006; Stocking & Lord, 1983). The question then arises as to what should be done about the anchor items that exhibit differential performance across samples of examinees taking different forms of an educational assessment. In many, if not most, assessment programs, outlier anchor items are removed from the anchor set before finalizing the equating. Different assessment programs use different methods and rules for determining whether the performance of an anchor item has changed sufficiently between assessment years (Gu, Lall, Monfils, & Jiang, 2010). However, as pointed out by Michaelides (2006), educational and accountability systems are designed to improved examinees' performance, and when items reflecting changes in achievement are removed on the basis of statistical criteria alone, the improvements brought about by the implementation of innovative programs, reallocation of resources, and other reform efforts may be adjusted away. In addition, the validity of the achievement test scores thus obtained may be affected as a result of such practices (Linn, 1990).

It is therefore prudent to follow the advice of Yen and Fitzpatrick (2006), who caution against removing anchor items for achievement tests if the change in performance from the previous use can be attributed to differential learning or minor sampling variation. However, items whose measurement properties have changed due to nonachievement factors or items with poorly estimated parameters warrant removal from the anchor set. Another consideration when removing anchor items is the effect on the content representativeness of the equating set. Klein and Jarjoura (1985) reported that poor content representation in the equating set may lead to substantial equating error. Similar concerns were expressed by Cook and Petersen (1987). In a recent simulation study, Wei (2010) found that violations of representativeness of items composing the anchor set can have a significant impact on the stability of scale scores across multiple years. Moreover, violations to content representativeness had a greater impact on score stability compared to violations of statistical representativeness.

Researchers have proposed various modifications to the mean difference equating method to reduce the need to delete outlier anchor items. Examples of the modifications include a weighted method where the weights are inversely proportional to the standard error of estimates of the item difficulty parameters (Linn, Levine, Hastings, & Wardrop, 1981) and a robust method that gives smaller weights to outlying points (Bejar & Wingersky, 1981).

Under the Rasch equating framework, removing outlier anchor items may have greater impact on calculation of the equating constant because means are sensitive to outliers, particularly when a small number of items is included in the anchor set. Two commonly used criteria for identifying anomalous performance in anchor items for potential removal from the anchor set include the 0.3 logit difference rule (G. E. Miller, Rotou, & Twing, 2004) and the robust z statistic (Hogg, 1979; Huynh, 1982). The information-weighted linking constant approach (Cohen, Jiang, & Yu, 2008), on the other hand, reduces the impact of outlier anchor items and thus mitigates the need to remove outliers by incorporating the uncertainty inherent in the item parameter estimates. Using 3 years of data (2005–2007) from the Ohio Achievement Tests, Cohen et al. (2008) found that the information-weighted mean difference method produced the smallest changes in year-to-year comparison of test scores when compared to the mean difference equating method, where all anchor items

are retained and where items are iteratively removed using the 0.3 logit rule. However, this study used real data, and thus the minimal changes observed do not indicate that one method is better than another simply because the true changes are unknown.

The purpose of this study is to further explore the performance of the information-weighted mean difference method when equating educational achievement tests with a NEAT design with a relatively long equating chain (5 years). Comparisons are made to the mean difference method without anchor item removal and the mean difference method with outlier removal using the 0.3 logit rule or robust *z* statistic. In addition, to fully evaluate the different approaches to equating, this study uses simulated data varying a number of factors, including sample size, distribution of test item difficulty, anchor test length, and percentage of anchor items displaying outlier behavior.

Methods

Data Generation

Simulated data were used in this study. An SAS program was written to generate the item response data under the Rasch model. With the Rasch model, the probability that a person with ability θ will respond correctly to item *i* is given by

$$P\left(U_i = 1|\theta\right) = P_i(\theta) = \frac{1}{1 + e^{-(\theta - b_i)}},$$

where U_i is the response to item *i*, 1 if correct and 0 if incorrect, and b_i is the threshold parameter of item *i*, characterizing its difficulty.

Data were generated to simulate five administrations of a typical mathematics examination consisting of 80 multiple-choice items. For each testing occasion, examinee ability was generated from a normal distribution with mean of 0 and variance of 1. The following factors were varied in the simulations:

1. *Sample size*. Two sample sizes (500 and 3,000 examinees) were examined. The 500-examinee condition represents a small sample size, and the 3,000-examinee condition represents a large sample size.
2. *Item difficulty distribution*. To model different conditions of misalignment of the distribution of item difficulty relative to the examinee ability distribution, true item difficulties (*b*-parameters) were simulated to produce test forms under two difficulty distribution conditions: approximately normal distribution drawn from *beta* (40, 40) and negatively skewed distribution drawn from *beta* (35, 5). The values were then rescaled so that the resulting item difficulties ranged from approximately −4 to 4. The normally distributed test forms had mean difficulty close to zero, and the negatively skewed test forms had mean difficulty close to 0.5. The negatively skewed distribution was included to model the condition of misalignment of the distribution of item difficulty to the examinee ability distribution.
3. *Anchor set*. A common guideline when constructing test forms is that the anchor set should have no fewer than 20 items or 25% of the number of items in the full test, whichever is larger (Angoff, 1971; Peterson, Marco, & Stewart, 1982; Wingersky, Cook, & Eignor, 1987). Following this guideline, an anchor set of 30 items was included as a study condition. Also included were anchor sets consisting of 15 items to represent those testing situations where the guidelines for minimum anchor size were not followed for reasons such as security or practice effect concerns.
4. *Outlier anchor items*. For each anchor set length, 0%, 20%, and 30% of the items were simulated to exhibit anomalous behavior. Specifically, these identified anchor items were randomly assigned to drift by ±0.3 logit with 60% chance of being assigned a −0.3 logit change. The 0.3 logit was chosen because item difficulty changes less than that were generally considered random and of no practical significance (Wright & Douglas, 1975).
5. *Equating constant calculation methods*. Four approaches to calculating the equating constant for use in mean equating were evaluated. The commonly used approach was based on new/reference form differences in mean anchor item difficulty without outlier removal (MD/None) and with the 0.3 logit rule (MD/0.3Logit) and robust *z* statistic rule (MD/RobustZ) to identify and remove outlier anchor items. In addition, the information-weighted mean difference procedure (MD/InfWt) was used. These methods are described in detail in a later section.

These factors were fully crossed, resulting in a total of 96 ($2 \times 2 \times 2 \times 3 \times 4$) conditions.

To generate item parameters for each of five forms (Forms A – E), item parameters were first generated for Form A, which functions as the base form administered in Year 1. Forms B, C, D, and E were administered in Years 2, 3, 4, and 5, respectively. Depending on the length of the anchor set, the item parameters for 15 or 30 of the items on Form B were obtained from Form A; the remaining item parameters for Form B were generated as described previously in Point 2. Item parameters for Forms C, D, and E were similarly generated with the parameters for the anchor set coming from the prior form. Under this design, no anchor items served as common items in three consecutive forms, for example, anchor items between A and B were different from the anchor items between B and C.

For each set of conditions, examinee item responses were generated for each of the five forms. The forms were calibrated and equated, and scoring tables were derived as described below. This process was repeated 50 times.

Calibration, Equating, and Number Correct-to-Theta Conversion

The simulated data sets for each replication of the five test forms were calibrated separately with the computer program WINSTEPS (Linacre, 2006) using the Rasch model (Rasch, 1960/1980). WINSTEPS by default set the origin of the scale to 0, and consequently, the item parameter estimates obtained for the base forms (forms administered in Year 1 and used to establish the score scale), particularly those simulated under the negatively skewed distribution, may not be on the same scale as the true parameters. To make the comparisons to the true parameters meaningful, the freely estimated parameters for the base forms (Form A) were first placed onto the same scale as the true parameters by setting the mean of observed item parameters to be the same as that of the true parameters. These Form A parameters were then considered to form the base scale to which the item parameters of the other four forms (Forms B – E) were equated through the anchor items in the previous form using each of the following approaches:

1. *Mean difference (MD/None)*. The equating constant is calculated as the average *b*-parameter of the anchor items on the reference form minus the average *b*-parameter of the same items on the new form. The equating constant is then added to the item difficulty parameters on the new form so this form is on the same scale as that established by the base form.
2. *Mean difference with the 0.3 logit rule (MD/0.3Logit)*. The equating constant is calculated with the mean difference method mentioned previously in Point 1 and is used to adjust for the item difficulties of the new form. However, an additional step — anchor stability analysis — is carried out. Specifically, the adjusted anchor item parameters on the new form are compared to the corresponding parameters on the reference form. Anchor items with absolute difference greater than 0.3 are removed from the anchor set, and the equating constant is recalculated with the remaining anchor items. This anchor stability analysis is repeated until no more anchor items are identified for removal. The equating constant calculated with this "stable" anchor set is used to place item parameters from the new form onto the scale of the base form.
3. *Mean difference with the Robust z statistic (MD/RobustZ)*. This procedure is similar to Point 2 described earlier in that it uses the mean difference method while also doing iterative anchor stability analyses. However, the robust *z* statistic, rather than the 0.3 logit rule, is used to identify anchor items that exhibit outlier behavior. Using the median (Mdn) and 0.74*interquartile range (IQR) to emulate the mean and standard deviation for a standard normal distribution, a robust *z* statistic (Hogg, 1979) can be written as

$$\text{Robust } z = \frac{d - \text{Mdn}\,(d)}{0.74\text{IQR}\,(d)},$$

where d is the difference between the Rasch difficulties of the new and reference forms for the items in the anchor set (i.e., $b_{iR} - b_{iN}$), Mdn(d) is the median of d, and IQR(d) is the interquartile range of d. As the robust z follows a standard normal distribution, the critical value of 1.645 was used to test for statistically significant differences between new and reference item difficulty values. That is, the item with the largest robust z value, if greater than 1.645, is removed from the anchor set. This process is repeated until no more anchor items are identified for removal. The equating constant calculated with the remaining anchor items is used to place item parameters from the new form onto the scale of the base form.

4 *Information-weighted mean difference (MD/InfWt)*. With the MD/InfWt procedure (Cohen et al., 2008), the equating constant used to place the new form (Form N) onto the scale of the reference form (Form R) is calculated as

$$d = \frac{1}{\sum_{i=1}^{I} \frac{1}{\sigma_{iR}^2 + \sigma_{iN}^2}} \sum_{i=1}^{I} \frac{b_{iR} - b_{iN}}{\sigma_{iR}^2 + \sigma_{iN}^2},$$

where b_{iR} and b_{iN} are the estimated item difficulty of the anchor item i in Forms R and N, respectively; σ_{iR} is the standard error of the b-parameter for item i in Form R; σ_{iN} is the standard error of the b-parameter for item i in Form N; and $1/(\sigma_{iR}^2 + \sigma_{iN}^2)$ is the information weight based on item calibration standard errors. The MD/InfWt method does not remove any outlier anchor items.

Once the equated item parameters were obtained for each form, the number-correct-score-to-theta relationship was obtained by summing over items the probability of correct response conditional on thetas and then finding the theta that corresponds to integer values of the summed probabilities (Lord, 1980). Two theta values 0.6 and 1.7 were arbitrarily chosen as the cut-scores to distinguish three performance levels on the base scale: not proficient, proficient, and advanced. For each form, the percentage of examinees classified at proficient and above and, at advanced performance levels, were calculated and compared across equating methods.

Evaluation Criteria

Several criteria were used to evaluate the performance of the four equating methods within the different conditions studied. The first criterion examined was the differences of the equating constants from the expected values. The second criterion examined the differences between estimated and true item parameters. Two indices, bias and the root mean square error (RMSE), were calculated as follows:

$$\text{BIAS}_b = \frac{\sum_{r=1}^{50} \sum_{i=1}^{80} \left(\widehat{b}_{ir} - b_i \right)}{80 * 50}$$

$$\text{RMSE}_b = \left[\frac{\sum_{r=1}^{50} \sum_{i=1}^{80} \left(\widehat{b}_{ir} - b_i \right)^2}{80 * 50} \right]^{1/2},$$

where b_i is the true item difficulty value for item i, $\widehat{b_i}$ is the estimated item difficulty for item i, and r is the number of replications. In the third method, differences between percentages of students classified into the proficient and advance performance category were calculated and compared for each of the four equating methods.

Results

Summary Statistics for Generated Data

Table 1 shows the average theta values generated for each of the 5 years across 50 replications for sample sizes of 500 and 3,000 examinees, respectively. These values were used in simulating item responses under each condition studied. Tables 2 and 3 list summary statistics for the true item difficulty values (generated b-parameters) for the total test and anchor sets for each form. These statistics represent the average over 50 replications. The item difficulty values used in the calculations of these statistics were prior to the introduction of changes in the anchor item difficulties to simulate outlier behavior. The results in Tables 2 and 3 show that within each condition, the five test forms were roughly parallel to each other in terms of difficulty. In addition, the anchor sets were similar in difficulty to the total test for each of the five forms.

Table 1 Summary Statistics for True Ability Distribution Across Replications

Sample size	Statistic	Year 1	Year 2	Year 3	Year 4	Year 5
500	Mean	0.0006	−0.0009	−0.0026	0.0063	−0.0027
	SD	0.9922	0.9934	0.9961	0.9945	1.0096
3,000	Mean	0.0055	−0.0010	−0.0018	−0.0022	−0.0012
	SD	1.0007	0.9997	0.9987	1.0001	1.0012

Table 2 Summary Statistics for True Form Item Difficulty With Anchor Set = 15 Items

		Year 1		Year 2		Year 3		Year 4		Year 5	
Condition	Statistic	Total test	Anchor test	Total test	Anchor test	Total test	Anchor test	Total test	Anchor test	Total test	Anchor test
0% Outliers											
b distribution, normal	Mean	0.0139	–	0.0093	−0.0334	−0.0004	0.0332	0.0048	0.0109	0.0391	0.0244
	SD	0.7942	–	0.7909	0.7663	0.8104	0.7884	0.8160	0.8022	0.8069	0.7867
	Skewness	0.0340	–	0.0329	0.0422	−0.0433	0.0211	−0.0162	−0.0255	−0.0954	−0.1532
b distribution, negatively skewed	Mean	0.5279	–	0.4925	0.4984	0.4970	0.4784	0.5102	0.4958	0.4846	0.5088
	SD	0.8016	–	0.8162	0.7804	0.8327	0.7976	0.8222	0.8121	0.8250	0.7906
	Skewness	−0.5939	–	−0.6500	−0.5039	−0.6657	−0.4640	−0.6539	−0.4305	−0.6417	−0.5831
20% Outliers											
b distribution, normal	Mean	0.0012	–	0.0051	0.0069	0.0051	−0.0311	0.0011	0.0209	0.0078	0.0231
	SD	0.8163	–	0.8100	0.8050	0.8105	0.8455	0.7993	0.7966	0.7876	0.7618
	Skewness	−0.0247	–	0.0432	−0.0080	0.0205	0.1574	−0.0194	0.1046	0.0022	0.0311
b distribution, negatively skewed	Mean	0.5165	–	0.5024	0.4910	0.4834	0.4788	0.4743	0.4963	0.4800	0.4218
	SD	0.8191	–	0.8249	0.8207	0.8254	0.8225	0.8404	0.8082	0.8363	0.8718
	Skewness	−0.7055	–	−0.6272	−0.6364	−0.6258	−0.5412	−0.6592	−0.4786	−0.6929	−0.6913
30% Outliers											
b distribution, normal	Mean	−0.0007	–	0.0046	0.0045	0.0134	0.0002	−0.0040	0.0142	−0.0203	−0.0117
	SD	0.8055	–	0.8054	0.7816	0.8107	0.7922	0.8171	0.8033	0.7971	0.8087
	Skewness	0.0235	–	0.0030	0.0056	−0.0010	0.0767	0.0186	0.0012	0.0315	−0.0737
b distribution, negatively skewed	Mean	0.4947	–	0.4842	0.4810	0.5287	0.5046	0.4933	0.5272	0.4923	0.4739
	SD	0.8275	–	0.8439	0.8369	0.8310	0.8451	0.8202	0.8008	0.8271	0.8362
	Skewness	−0.7021	–	−0.6001	−0.6423	−0.6521	−0.4823	−0.6402	−0.4151	−0.5972	−0.4836

Bias

Figure 1 shows bias for the item difficulty estimates averaged over the 50 replications under each simulated condition. It should be noted that the true item difficulty values for the anchor items (i.e., difficulty values of the anchor items in the reference form) were used in the calculation of bias. These results show that when there were no outlier items present in the anchor set, bias was relatively close to 0 across all four equating methods. Also, bias was the smallest with a sample size of 3,000 examinees and with an anchor set of 30 items.

When outliers were present, bias was similar for the MD/None and MD/InfWt methods under all conditions. Furthermore, bias for these two equating methods tended to be larger compared to the MD/0.3Logit and MD/RobustZ methods. Some exceptions to this trend for the MD/0.3Logit method were found, however, under conditions with 3,000 examinees, where this method tended to produce the largest bias. Under the same condition (3,000 examinees), the MD/RobustZ method had the smallest bias, which was similar to that under the no outlier condition. Bias for the MD/0.3Logit method varied depending on the sample size and percentage of outlier anchor items. With a sample size of 500 examinees, bias for the MD/0.3Logit method was generally between the bias for the MD/None and MD/InfWt methods and the MD/RobustZ method. With a sample size of 3,000 examinees and with 20% outlier anchor items, bias was close to those of the MD/None and MD/InfWt methods. With 30% outlier anchor items (and sample size of 3,000 examinees), bias was generally larger compared to those from the other equating methods. The bias for all methods tended to increase from Year 1 to Year 5. This pattern was consistent across sample size, number of anchor items, and distribution of item difficulty. This pattern was also evident at the different levels for the percentage of outlier anchor items, although the pattern was less evident with 0% outliers at the larger sample size and anchor set lengths.

Table 3 Summary Statistics for True Form Item Difficulty With Anchor Set = 30 Items

		Year 1		Year 2		Year 3		Year 4		Year 5	
Condition	Statistic	Total test	Anchor test	Total test	Anchor test	Total test	Anchor test	Total test	Anchor test	Total test	Anchor test
0% Outliers											
b distribution, normal	Mean	−0.0002	–	−0.0032	−0.0006	−0.0078	−0.0158	0.0000	0.0023	−0.0017	0.0115
	SD	0.7840	–	0.8037	0.7853	0.7960	0.8042	0.7928	0.7823	0.7964	0.7933
	Skewness	0.0582	–	0.0212	0.1217	0.0030	0.0041	0.0637	0.0287	0.0262	0.0387
b distribution, negatively skewed	Mean	0.4990	–	0.4948	0.4897	0.5027	0.5135	0.5296	0.4987	0.5391	0.5702
	SD	0.8296	–	0.8152	0.8072	0.8258	0.7965	0.8016	0.8312	0.7871	0.7579
	Skewness	−0.6816	–	−0.5891	−0.5665	−0.6396	−0.5032	−0.6417	−0.6990	−0.5442	−0.4620
20% Outliers											
b distribution, normal	Mean	−0.0006	–	−0.0147	0.0002	−0.0267	−0.0242	−0.0001	−0.0175	0.0063	−0.0080
	SD	0.8157	–	0.8015	0.8163	0.7922	0.7776	0.8025	0.7934	0.8042	0.7942
	Skewness	−0.0194	–	0.0097	−0.0125	0.0187	−0.0373	0.0346	0.0101	0.0022	0.0544
b distribution, negatively skewed	Mean	0.4848	–	0.4913	0.4797	0.4977	0.5197	0.4880	0.4885	0.5057	0.5129
	SD	0.8262	–	0.8395	0.8274	0.8275	0.8333	0.8193	0.8096	0.8348	0.7989
	Skewness	−0.5770	–	−0.6342	−0.5302	−0.6329	−0.6898	−0.6549	−0.5440	−0.7246	−0.6774
30% Outliers											
b distribution, normal	Mean	0.0095	–	0.0024	0.0264	0.0125	0.0132	−0.0004	0.0051	0.0141	0.0132
	SD	0.7828	–	0.8102	0.7766	0.8136	0.8091	0.8010	0.7876	0.8103	0.8057
	Skewness	0.0151	–	0.0523	−0.0117	−0.0116	0.0091	−0.0204	0.0394	0.0052	−0.0640
b distribution, negatively skewed	Mean	0.5180	–	0.5020	0.5059	0.5067	0.5249	0.4878	0.5146	0.4868	0.4712
	SD	0.8019	–	0.8114	0.7856	0.8237	0.8209	0.8380	0.8181	0.8287	0.8267
	Skewness	−0.5717	–	−0.5284	−0.5429	−0.6186	−0.5660	−0.6650	−0.6284	−0.6332	−0.5569

Root Mean Square Error

Figure 2 presents the RMSE under each simulated condition. Similar to the calculation of bias, the true item difficulty values for the anchor items (i.e., difficulty values of the anchor items in the reference form) were used in the RMSE calculation. These results show that, generally, the RMSE for MD/None and MD/InfWt was similar across sample size, anchor set length, percentage of outlier anchor items, and distribution of item difficulty. With a sample size of 500 examinees, the RMSE tended to be similar across the four equating methods, although small differences were observed in Years 4 and 5 among them. This pattern was consistent across anchor set length, percentage of outliers, and distribution of item difficulty values.

With a sample size of 3,000 and no outlier anchor items, the RMSEs were similar among all of the equating methods. However, when outlier anchor items were present, the differences in RMSE varied more by method as percentage of outliers increased. The MD/RobustZ method, on average, produced the smallest RMSE compared to the other three equating methods. Also, with 20% outliers and sample size of 3,000 examinees, RMSE was similar across the MD/None, MD/InfWt, and MD/0.3Logit methods. This pattern was consistent across anchor set length. With 30% outlier anchor items and sample size of 3,000 examinees, the MD/0.3Logit method had the largest RMSE, followed by MD/None and MD/InfWt (both of which performed similarly) and the MD/RobustZ method, which had the least RMSE. RMSE for all methods increased from Year 1 to Year 5. This pattern was consistent across sample size, number of anchor items, percentage of outliers, and distribution of item difficulty. Note, however, that for the MD/RobustZ method with a sample size of 3,000 examinees, the increase in RMSE from Year 2 to Year 5 was very small across most conditions, with the exception of 30% outliers and an anchor set of 15 items.

Equating Constants

The averages of the equating constants across the 50 replications are shown graphically in Figure 3 for each of the four equating methods. The equating constants were close to 0 when the item difficulty values were normally distributed and close to 0.5 when the distribution was negatively skewed. The sizable nonzero constants for the negatively skewed condition are expected given that the difficulty of the skewed reference form was first equated back to the scale of the true parameters and the difficulty of items in the new forms (centered at 0) need to be adjusted to the same scale as the base form

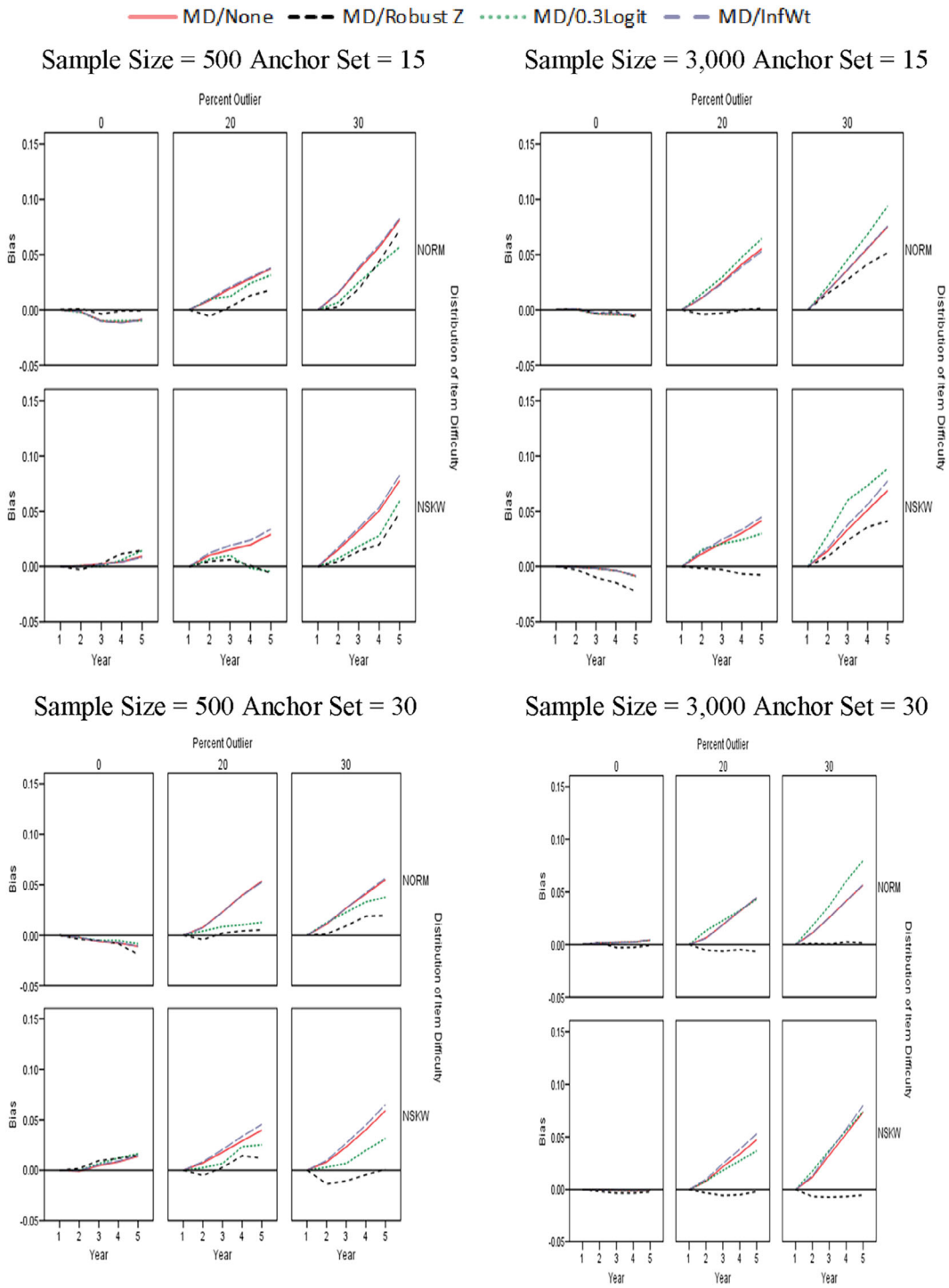

Figure 1 Item difficulty bias.

parameters. The results indicate that with 0% outliers, the four equating methods had similar equating constants across all conditions. With a sample size of 500 examinees, the MD/None and MD/InfWt methods had more similar values for the equating constants, while MD/RobustZ and MD/0.3Logit values were more similar, although overall, the differences in equating constants were quite small. With a sample size of 3,000 examinees, MD/None, MD/InfWt, and MD/0.3Logit had similar equating constants that were slightly different from the MD/RobustZ method. These trends were consistent across anchor set length and distribution of item difficulty values. Similar to the increasing trend observed for bias and RMSE,

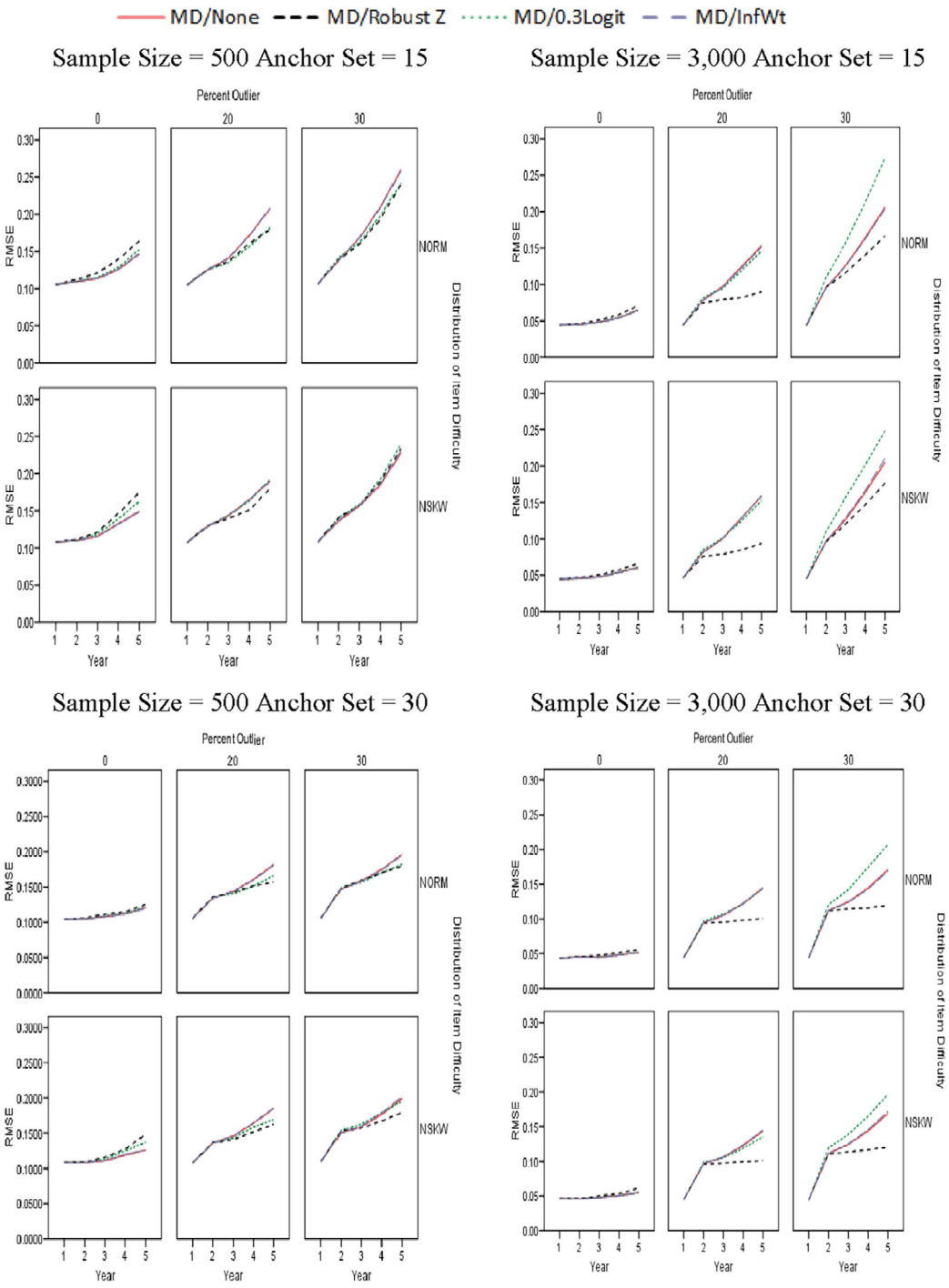

Figure 2 Root mean square error of the item difficulty.

the equating constants for the different equating methods increase slightly from Year 2 to Year 5, the increase being more noticeable at Years 4 and 5 for the MD/None, MD/0.3Logit, and MD/InfWt methods.

Percentage of Anchor Items Removed

Figure 4 shows the percentage of anchor items removed by the MD/RobustZ and MD/0.3Logit methods. Overall, the MD/RobustZ method removed a greater percentage of anchor items than the MD/0.3Logit method did. When there

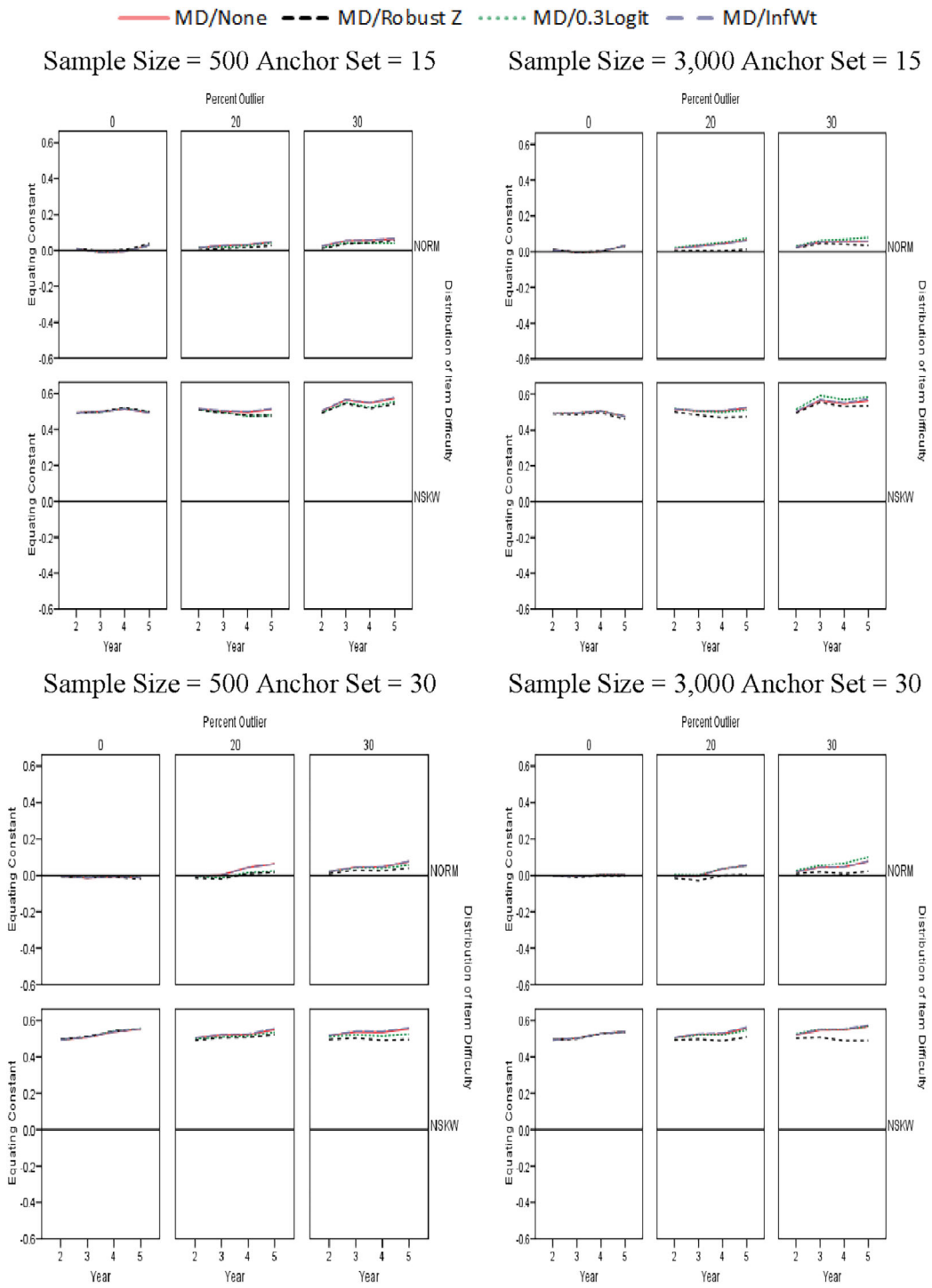

Figure 3 Equating constants.

were no outliers, the MD/RobustZ method removed items from the anchor set for all conditions (14–25% anchor items removed). In contrast, when no outlier anchor items were present, the MD/0.3Logit method removed items (0–5%) from the anchor set only when the sample size was 500 examinees. When outlier anchor items were present, slightly more items were removed by the MD/RobustZ method when sample size was 3,000 examinees compared to when sample size was 500 examinees. This is in contrast to the MD/0.3Logit method, where slightly more items were removed with a sample size of 500 examinees compared to a sample size of 3,000 examinees. This pattern was consistent across anchor set length

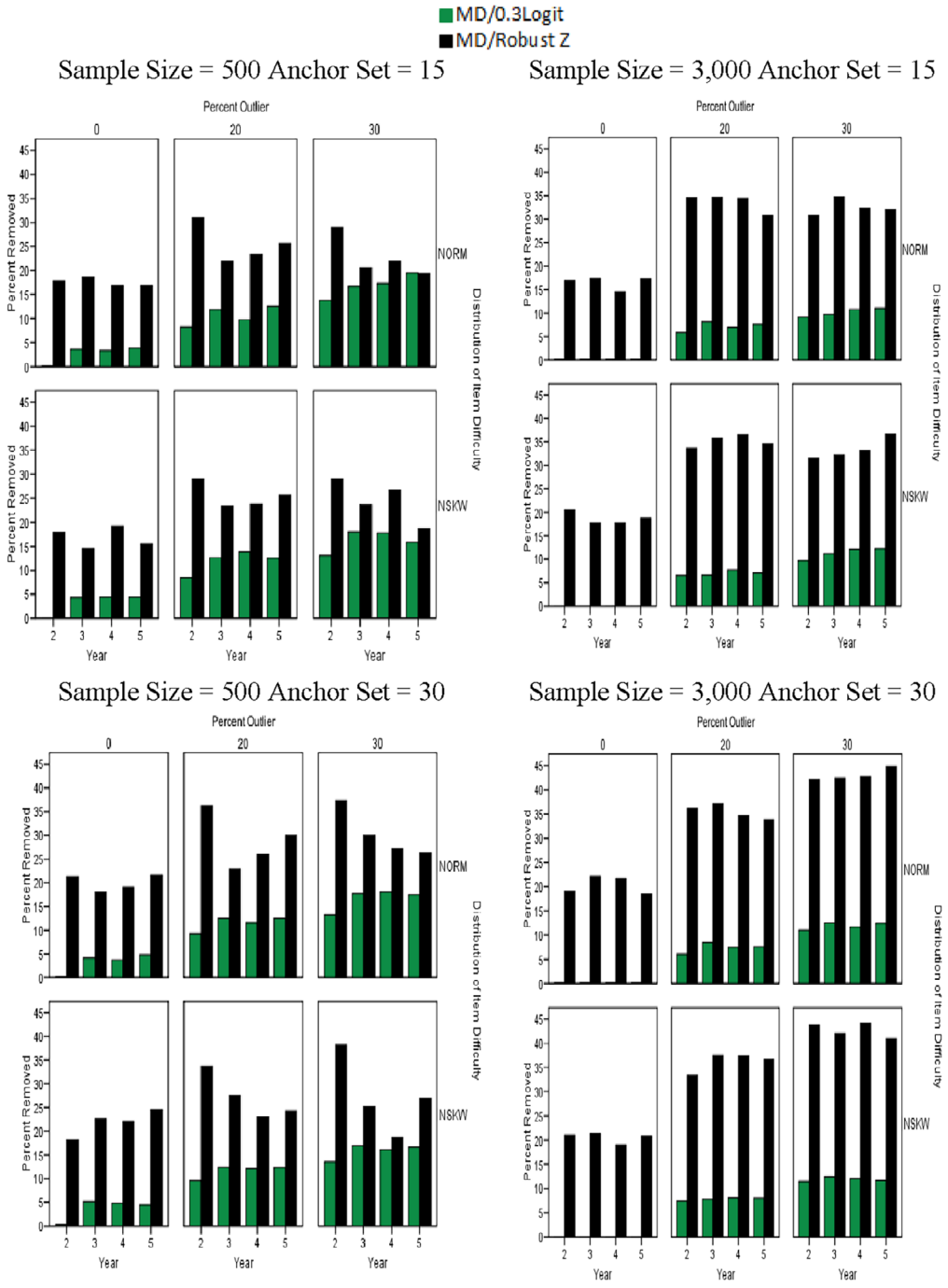

Figure 4 Percentage of anchor items removed by MD/RobustZ and MD/0.3Logit.

and distribution of item difficulty. Overall, the MD/RobustZ method removed greater percentages of anchor items than the actual percentages of generated outlier anchor items.

Performance Level Classification

For each condition, after application of the equating constants, true-score-to-theta conversion tables were produced for each form and the performance classification rates were obtained for samples in the corresponding conditions. Performance classifications across years were expected to be close to the percentages in Year 1. Figure 5 shows the

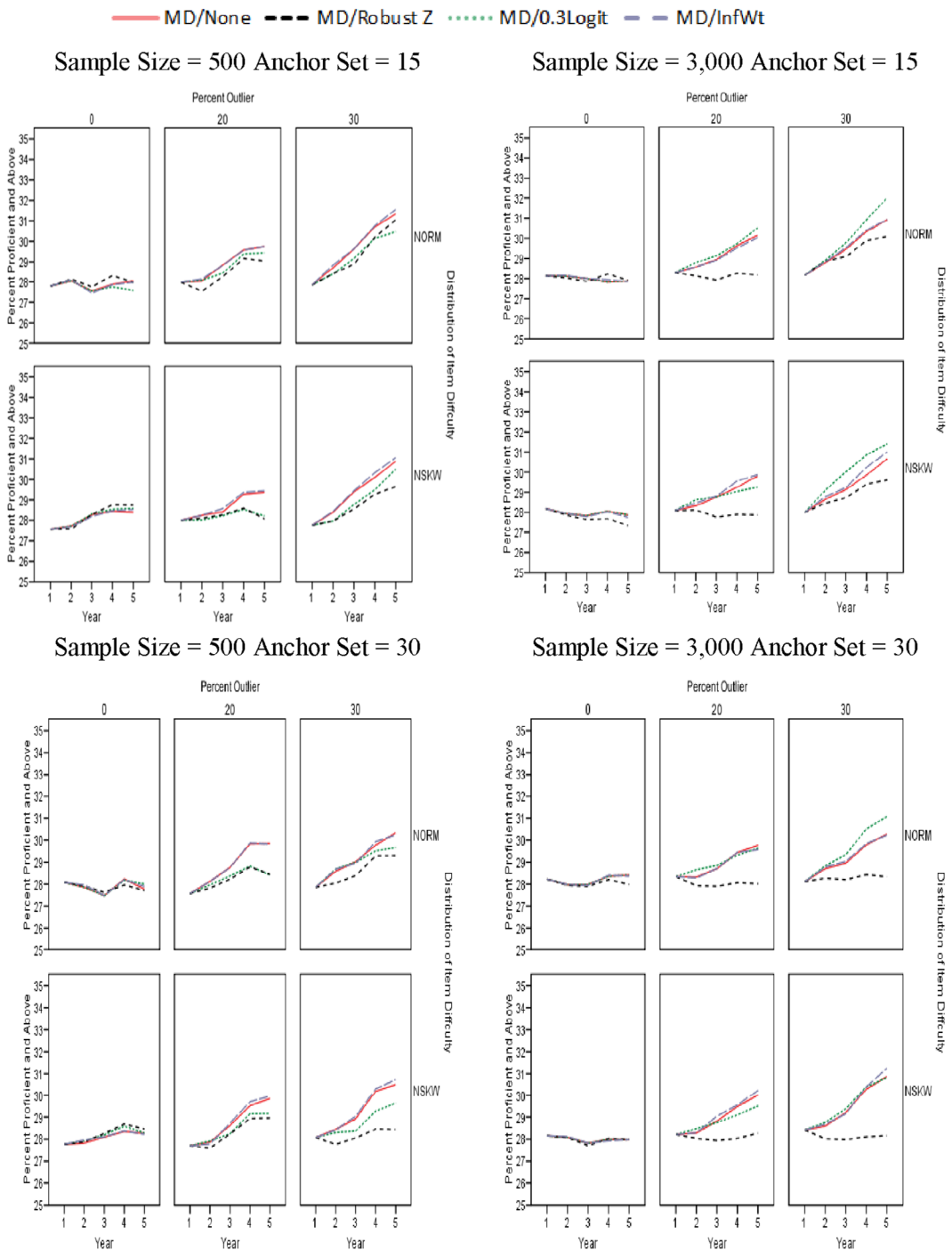

Figure 5 Percentage at proficient and above classification.

percentages of examinees classified into the proficient and above performance categories, while Figure 6 shows the percentage classified into the advanced category. Similar trends were observed for examinee classifications into the two performance levels. Results show that in conditions without outlier anchor items, performance classifications were similar across the four equating methods. When outlier anchor items were present, the MD/None and MD/InfWt methods produced similar percentages of examinees classified into each of the performance levels. This trend was consistent across all conditions. In addition, the percentage classified generally increased from Year 1 to Year 5. Of the four methods, the MD/RobustZ method tended to have the most stable percentage classifications across years. This trend is particularly

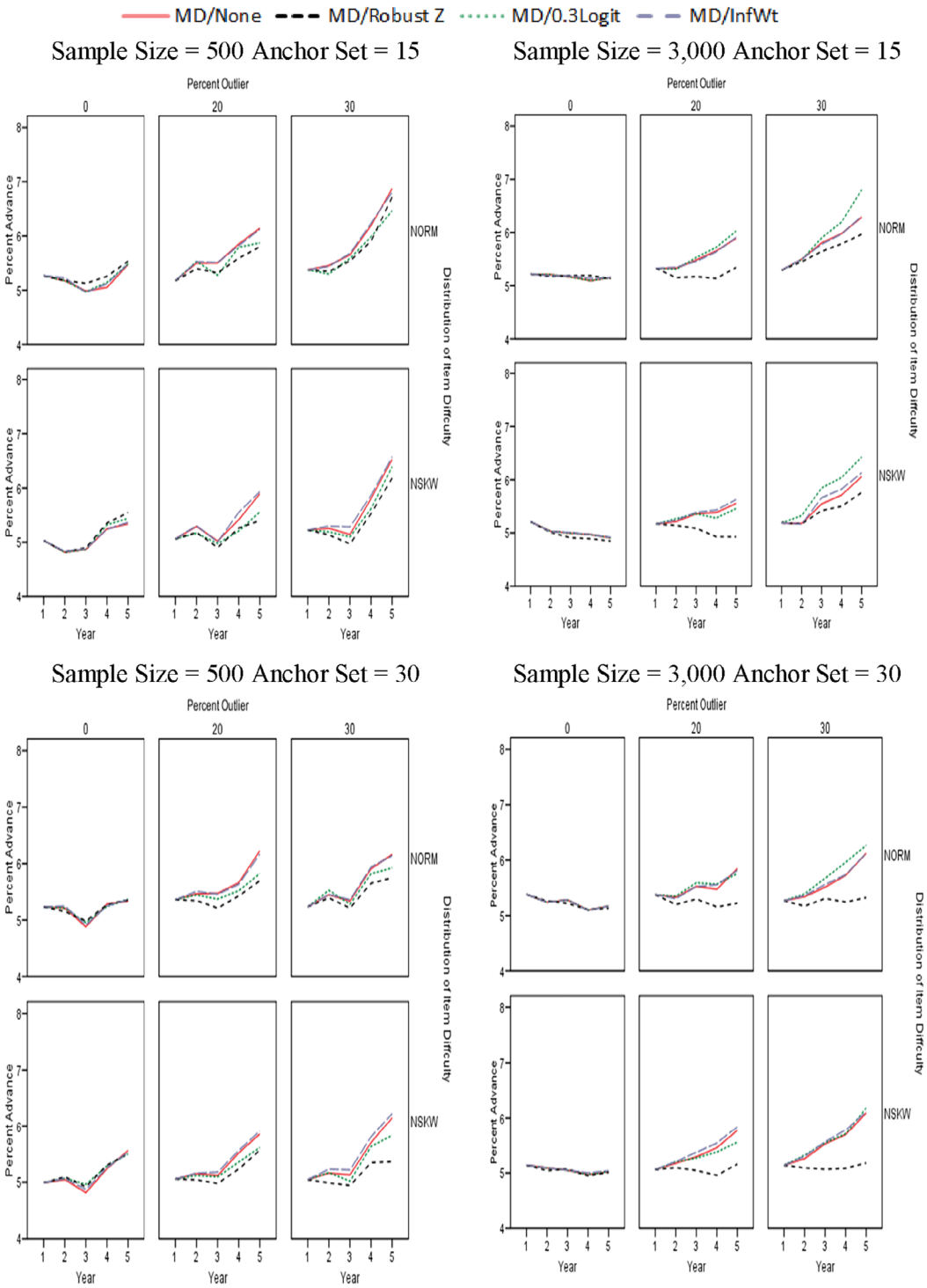

Figure 6 Percentage advanced classification.

evident at a sample size of 3,000 examinees. The exception to this trend is with 15 anchor items and 30% outliers, where there was a noticeable increase in the percentages of examinees classified at the proficient and advanced categories from Year 1 to Year 5. The other three methods showed similar increases from Years 1 to 5. With a sample size of 500, the increases in performance classification across years using the MD/0.3Logit and MD/RobustZ methods were slightly more variable.

Discussion

This study examined four approaches to calculating the equating constants under the Rasch framework—mean difference without outlier removal, mean difference with outlier removal using the 0.3 logit rule or robust *z* statistic, and the information-weighted mean difference methods—and the subsequent impact on equating results, including the consistency of scores across 5 years of simulated data. Factors manipulated were sample size, anchor test length, percentage of anchor items displaying outlier behavior, and distribution of item difficulty relative to examinee ability. The results indicate that, in general, the mean difference without outlier removal and information-weighted mean difference methods perform similarly across all conditions. In addition, with larger sample size, the mean difference with outlier removal using 0.3 logit rule method performed similarly to these two methods. The mean difference with outlier removal using robust *z* performed most differently from the other three methods of calculating the equating constant. This method removed a larger percentage of the anchor items compared to the mean difference with 0.3 logit method but seemed to produce the most stable trend in performance classification across the 5 years, particularly at the larger sample size.

One of the key findings is that methods that retained all items in the anchor set performed similarly under most conditions. Another key finding is that the MD/RobustZ method overestimates the number of outlier anchor items. This finding is in contrast to the MD/0.3Logit method, which underestimates the number of true outliers. These results are consistent with previous research (Huynh & Rawls, 2011; Murphy, Little, Fan, Lin, & Kirkpatrick, 2010). One consideration is to use a more stringent criterion for determining statistical significance with MD/RobustZ, as suggested by other researchers (Agresti & Finlay, 2009; Huynh & Meyer, 2010). In addition, results of this study support previous research that suggested that more accurate and more stable estimates are obtained with slightly larger anchor sets and larger calibration sample sizes. Not surprisingly, this study also found that lower percentages of outliers in the anchor set will produce more accurate and stable estimates.

However, removal of items with seemingly anomalous behavior, with decisions based on statistical criteria alone, is not recommended. The anchor item set used in equating provides the foundation on which annual comparisons are made. To maintain stability, it is important to maintain the statistical and content representativeness of the anchor set. The removal of items from the anchor set due to anomalous behavior can affect the stability of equating results as well as the validity of the test scores. Change in anchor difficulty may reflect real differences in examinee performance due to growth, changes in curricular emphasis, and other achievement-related factors. Therefore it is advisable to follow the recommendation of Yen and Fitzpatrick (2006) only to remove items from the anchor items if the differences in performance can be attributed to construct-irrelevant factors like context effects, poor fit, and overexposure. To this end, it is critical to establish rigorous guidelines for selecting anchor items and a priori rules for identifying and removing anchor items. It is also important to include procedures to monitor the performance of items over time and include routine checks for scale drift in operational testing programs.

On a final note, as with any simulation study, caution should be used in interpreting results beyond the parameters of the study.

References

Agresti, A., & Finlay, B. (2009). *Statistical methods for the social sciences* (4th ed.). Upper Saddle River, NJ: Pearson Prentice Hall.

Angoff, W. H. (1971). Scales, norms and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508–600). Washington, DC: American Council on Education.

Bejar, I. I., & Wingersky, M. S. (1981). *An application of item response theory to equating the Test of Standard Written English* (College Board Report No. CBR-81-08). Princeton NJ: Educational Testing Service.

Bock, R. D., Muraki, E., & Pfeiffenberger, W. (1988). Item pool maintenance in the presence of item parameter drift. *Journal of Educational Measurement*, *25*, 275–285. https://doi.org/10.1111/j.1745-3984.1988.tb00308.x

Cassels, J. R. T., & Johnstone, A. H. (1984). The effect of language on student performance on multiple choice tests in chemistry. *Journal of Chemical Education*, *61*, 613–615. https://doi.org/10.1021/ed061p613

Cizek, G. J. (1994). The effect of altering the position of options in a multiple-choice examination. *Educational and Psychological Measurement*, *54*, 8–20. https://doi.org/10.1177/0013164494054001002

Cohen, J., Jiang T., & Yu, P. (2008) *Information-weighted linking constants*. Washington, DC: American Institute for Research.

Cook, L. L., Eignor, D. R., & Taft, H. L. (1998). A comparative study of the effects of recency of instruction on the stability of IRT and conventional item parameter estimates. *Journal of Educational Measurement*, *25*, 31–45. https://doi.org/10.1111/j.1745-3984.1988.tb00289.x

Cook, L. L., & Petersen, N. S. (1987). Problems related to the use of conventional and item response theory equating methods in less than optimal circumstances. *Applied Psychological Measurement*, *11*, 225–244. https://doi.org/10.1177/014662168701100302

Dorans, N. J. (1986). The impact of item deletion on equating conversions and reported score distributions. *Journal of Educational Measurement, 25*, 245–264. https://doi.org/10.1111/j.1745-3984.1986.tb00250.x

Eignor, D. R., & Cook, L. L. (1983, April). *An investigation of the feasibility of using item response theory in the preequating of aptitude tests*. Paper presented at the annual meeting of the American Educational Research Association, Montreal, QC.

Fitzpatrick, A. R., Ercikan, K., Yen, W. M., & Ferrara, S. (1998). The consistency between raters scoring in different test years. *Applied Measurement in Education*, *11*, 195–208. https://doi.org/10.1207/s15324818ame1102_5

Gu, L., Lall, V. F., Monfils, L., & Jiang, Y. (2010, April). *Evaluating anchor items for outliers in IRT common item equating: A review of commonly used methods and flagging criteri*a. Paper presented at the annual meeting of the National Council on Measurement in Education, Denver, CO.

Hanson, B. A., & Feinstein, Z. S. (1997). *Application of a polynomial log linear model to assessing differential item function for common items in the common-item equating design* (ACT Research Report Series No. 97-1). Iowa City, IA: ACT Inc.

Hogg, R. V. (1979). Statistical robustness: One view on its use in applications today. *The American Statistician*, *33*, 108–115.

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 187–220). Westport, CT: Praeger.

Hu, H., Rogers, W. T., & Vukmirovic, Z. (2008). Investigation of IRT-based equating methods in the presence of outlier common items. *Applied Psychological Measurement*, *32*, 311–333. https://doi.org/10.1177/0146621606292215

Huynh, H. (1982). A comparison of four approaches to robust regression. *Psychological Bulletin*, *92*, 505–512. https://doi.org/10.1037/0033-2909.92.2.505

Huynh, H., & Meyer, P. (2010). Use of robust *z* in detecting unstable items in item response theory models. *Practical Assessment, Research and Evaluation*, *15*(2). Retrieved from http://pareonline.net/getvn.asp?v=15&n=12

Huynh, H., & Rawls, A. (2011). A comparison between robust *z* and 0.3-logit difference procedures in assessing stability of linking items for the Rasch model. *Journal of Applied Measurement, 12*(2), 96–105.

Karkee, T., & Choi, S., (2005, April). *Impact of eliminating anchor items flagged from statistical criteria on test score classifications in common item equating*. Paper presented at the annual meeting of the American Educational Research Association, Montreal, QC.

Kingston, N. M., & Dorans, N. J. (1984). Item location effects and their implications for IRT equating and adaptive testing. *Applied Psychological Measurement*, *8*, 147–154. https://doi.org/10.1177/014662168400800202

Klein, L. W., & Jarjoura, D. (1985). The importance of content representation for common-item equating with nonrandom groups. *Journal of Educational Measurement*, *22*, 197–206. https://doi.org/10.1111/j.1745-3984.1985.tb01058.x

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York, NY: Springer. https://doi.org/10.1007/978-1-4757-4310-4

Linacre, J. M. (2006). *WINSTEPS Rasch measurement computer program*. Chicago, IL: Winsteps.

Linn, R. L. (1990). Has item response theory increased the validity of achievement test scores? *Applied Measurement in Education*, *3*, 115–141. https://doi.org/10.1207/s15324818ame0302_1

Linn, R. L., Levine, M. V., Hastings, C. N., & Wardrop, J. L. (1981). Item bias in a test of reading comprehension. *Applied Psychological Measurement*, *5*, 159–173. https://doi.org/10.1177/014662168100500202

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

Masters, G. N. (1988). Item discrimination: When more is worse. *Journal of Educational Measurement*, *25*, 15–29. https://doi.org/10.1111/j.1745-3984.1988.tb00288.x

Michaelides, M. P. (2006). *Effects of misbehaving common items on aggregate scores and an application of the Mantel–Haenszel statistic in test equating*. Los Angeles, CA: Center for the Study of Evaluation, University of California, Los Angeles.

Miller, A. D., & Linn, R. L. (1988). Invariance of item characteristic functions with variations in instructional coverage. *Journal of Educational Measurement*, *25*, 205–219. https://doi.org/10.1111/j.1745-3984.1988.tb00303.x

Miller, G. E., Rotou, O., & Twing, J. S. (2004). Evaluation of the 0.3 logit screening criterion in common item equating. *Journal of Applied Measurement*, *5*, 172–177.

Murphy, S., Little, I., Fan, M., Lin, C., & Kirkpatrick, R. (2010, April). *The impact of different anchor stability methods on equating results and student performance*. Paper presented at the annual meeting of the National Council on Measurement in Education, Denver, CO.

Peterson, N. S., Marco, G. L., & Stewart, E. E. (1982). A test of the adequacy of linear equating methods. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 71–135). New York, NY: Academic Press.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago, IL: University of Chicago Press. (Original work published 1960)

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, *7*, 201–210. https://doi.org/10.1177/014662168300700208

Tollefson, N. (1987). A comparison of the item difficulty and item discrimination of multiple-choice items using the "none of the above" and one correct response options. *Educational and Psychological Measurement*, *47*, 377–383. https://doi.org/10.1177/0013164487472010

Wainer, H. (1999). Comparing the incomparable: An essay on the importance of big assumptions and scant evidence. *Educational Measurement: Issues and Practice*, *18*, 10–16. https://doi.org/10.1111/j.1745-3992.1999.tb00277.x

Wei, H. (2010). *Impact of non-representative anchor items on scale stability*. Paper presented at the annual meeting of the National Council on Measurement in Education, Denver, CO.

Wingersky, M. S., Cook, L. L., & Eignor, D. R. (1987). *Specifying the characteristics of linking items used for item response theory calibration* (Research Report No. RR-87-24). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1987.tb00228.x

Wright, B. D., & Douglas, G. A. (1975). *Best test design and self-tailored testing* (Research Memorandum No. 19). Chicago, IL: University of Chicago, Department of Education, Statistical Laboratory.

Yen, W. M. (1980). The extent, causes and importance of context effects on item parameters for two latent trait models. *Journal of Educational Measurement*, *17*, 297–311. https://doi.org/10.1111/j.1745-3984.1980.tb00833.x

Yen, W. M., & Fitzpatrick, A. (2006). Item response theory. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 111–153). Westport, CT: Praeger.

Suggested citation:

Manna, V. F., & Gu, L. (2019). *Different methods of adjusting for form difficulty under the Rasch model: Impact on consistency of assessment results* (Research Report No. RR-19-08). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12244

Action Editor: Gautam Puhan

Reviewers: Longjuan Liang and Lora Monfils